BRAZILIAN JOURNAL OF FOOD TECHNOLOGY

ISSN 1981-6723
Online version

ORIGINAL ARTICLE

Detecção de valores aberrantes em alguns experimentos interlaboratoriais para alimentos

Detecting outliers in some food interlaboratory experiments revisited

Elisabeth Borges Gonçalves[1]*, Ana Paula Guedes Alves[2], Paula Alves Martins[2]

[1]Embrapa Agroindústria de Alimentos, Rio de Janeiro/RJ - Brasil
[2]Universidade Estadual do Rio de Janeiro (UERJ), Instituto de Matemática e Estatística (IME), Rio de Janeiro/RJ - Brasil

***Corresponding Author:** Elisabeth Borges Gonçalves, Embrapa Agroindústria de Alimentos, Av. Américas, 29501, Guaratiba, CEP: 23020-470, Rio de Janeiro/RJ - Brasil, e-mail: elisabeth.goncalves@embrapa.br

Cite as: Gonçalves, E. B., Alves, A. P. G., & Martins, P. A. (2021). Detecting outliers in some food interlaboratory experiments revisited. *Brazilian Journal of Food Technology*, *24*, e2020273. https://doi.org/10.1590/1981-6723.27320

Resumo

Valores aberrantes são grande preocupação em laboratórios de alimentos, até pela heterogeneidade de muitas matrizes, e, embora haja diversos testes de hipóteses para identificá-los, faltam informações detalhadas sobre seus resultados. Gráficos, como desenhos esquemáticos, ramo-e-folhas e outros, não têm sido vistos em laboratórios com esse objetivo. Interlaboratoriais da literatura em alimentos foram revistos e tais gráficos foram empregados junto aos testes de Dixon, Grubbs e Hampel, além do escore Z_r. Pôde-se notar como os resultados desses métodos podem apresentar discordâncias, tanto entre os gráficos como entre os resultados numéricos, e entre os últimos e os primeiros. Dentre os gráficos empregados, além dos desenhos esquemáticos, os ramo-e-folhas trouxeram boa visualização de valores aberrantes. Nas técnicas numéricas, o escore Z_r apontou o maior número de valores aberrantes potenciais e o teste de Hampel se mostrou mais adequado. Houve conformidade entre os resultados desta pesquisa e os de um dos quatro interlaboratoriais revisitados, enquanto foram observadas divergências nos demais. Sugere-se o emprego, quando possível, dos modelos de delineamentos ao estudar os experimentos e atenção aos casos com mais de 1% de valores aberrantes.

Palavras-chave: Metrologia; Qualidade para laboratórios; Validação de métodos analíticos; Testes de proficiência; Estudos colaborativos; Desenho esquemático; Diagrama de ramo-e-folhas.

Abstract

Outliers are always frequent in all analytical laboratories and ever source of trouble, mainly in food owing to the heterogeneity in the matrixes. Although there are several formal statistical tests to identify those data, publications lack enough information about their practical results. Box plots, stem-and-leaf and other graphics are not included in a routine of these laboratories and their results are not known, therefore these and other graphics, in addition to Dixon, Grubbs and Hampel tests, being the last one more recent than the other, and the robust Z-score (Z_r) were employed when revisiting previously published interlaboratory experiments. The results showed some disagreement among the conclusions from the graphics, the same among the numerical methods and among graphics and numeric methods. In addition to the box

plots, the stem-and-leaf could provide an effective visualization of outliers. The Z_r score showed the highest number of potential aberrant values and the Hampel test was the most appropriated test. The results of outliers in one of the four revisited experiments were in agreement with the results of this research, however, those regarding the other three revisited interlaboratories were not. This research suggested to use the experimental design model to study the measures, when possible, and to examine the cases with more than 1% potential outliers.

Keywords: Metrology; Quality in laboratories; Single validation; Proficiency tests; Collaborative studies; Box plot; Stem-and-leaf.

1 Introdução

Valores aberrantes (*outliers)* perturbam pesquisadores há séculos e são rotina nos laboratórios de análises de alimentos em que muitas matrizes são heterogêneas. A preocupação permanece atual e efervescente em validação de métodos analíticos e bioanalíticos, e no cotidiano dos laboratórios.

A detecção desses valores é essencial em medidas espectroscópicas e difícil em espaços multidimensionais, nos quais começa a despontar, como Pang & Cao (2020). Infortunadamente, métodos multidimensionais (multivariados) têm falhado ao identificar valores aberrantes unidimensionais (univariados).

A noção desses valores é difusa, pois não são rigorosamente definidos, visto que seriam "dados que diferem bastante dos demais em conjunto", segundo Dunn & Clark (1987). A *Food and Drug Administration* (2006) os admite como dados "raramente" obtidos, "marcadamente diferentes" dos outros em séries de medições com métodos validados, e considera crítico para os laboratórios prover a gestão da qualidade de todos os resultados para suas disposições finais. A *Organisation for Economic Co-operation and Development* (2018) admite uma definição similar a essa, porém em uma amostra aleatória.

Segundo Molenaar et al. (2018), não seria fácil discriminar esses valores dos demais porque sua "diferença" não é bem definida. Seaman & Allen (2010) admitem que removê-los pode resultar perda de detalhes, ter grandes efeitos em análises, e que há grande dificuldade na presença de dois ou mais valores desse tipo.

A *Organisation for Economic Co-operation and Development* (2018), ao avaliar o potencial de sensibilização (SI) da pele de um produto químico em teste, orienta que se calcule esse potencial com e sem esse tipo de valor. Assim, esses valores podem interferir em um resultado de teste de um produto na pele.

Há diversos testes de hipóteses usados em laboratórios de alimentos para esses valores, mas seus resultados não são discutidos. São muito encontrados em: a. validação intralaboratorial (Horwitz, 1995; *International Organization for Standardization*, 2005); b. experimentos de precisão ou estudos colaborativos (Instituto Nacional de Metrologia, Qualidade e Tecnologia, 2016); c. ensaios de proficiência (*National Institute of Standards and Technology*, 2020), e também, segundo os últimos, d. em outros experimentos.

Nessa área, esses valores têm sido identificados através dos testes de Cochran (1941), Dixon (1950) e Grubbs (1969), *Association of Official Analytical Chemists* (1995, 2002), Instituto Nacional de Metrologia, Qualidade e Tecnologia (2016).

Segundo ambas (*Association of Official Analytical Chemists*, 1995, 2002), o máximo número de valores aberrantes (*outliers*) a serem descartados sem estudar sua origem seria de de dois em nove estudos (22,22%).

Peirce (1852) desenvolveu o primeiro teste para esses valores, que identifica mais de um valor aberrante em distribuição normal. O teste de Chauvenet (1863) admite P[(valor aberrante – média) > 2 × desvio padrão] ≈ 0,05 e também foi desenvolvido para normalidade. Esses dois testes são pouco aplicáveis a pequenos conjuntos de dados.

O teste de Cochran (1941), não paramétrico, é definido por (Equação 1):

$$C = \max s_i^2 / \sum s_i^2 \quad (1)$$

Em que s_i^2 é a variância estimada do i-ésimo laboratório, i=1, 2, ..., n. Pela definição, o teste se baseia em variâncias e também costuma ser empregado ao verificar homocedasticia.

Dixon (1950) estabeleceu Q como estatística de teste para valor aberrante, com $Q = |(X_n - X_{n-1})/(X_n - X_1)|$, *Xn* o último valor após ordenação e candidato a aberrante, quando: $3 \leq n \leq 7$; ou $Q = |(X_n - X_{n-1})/(X_n - X_1)|$, quando $8 \leq n \leq 10$; ou ainda $Q = |(X_n - X_{n-1})/(X_n - X_2)|$, para $11 \leq n \leq 13$, em dados com distribuição normal. Os *Pharmceutical Technology Editors* (2006) o apresentam e chegam a mencionar que esse teste não teria requerimentos distribucionais.

Esse teste é tão popular que Zhongya et al. (2020) realizaram simulações buscando ampliar seus limites para aplicações em séries temporais (previsões), aproveitanto o desenvolvimento de algoritmos atuais bastante velozes.

O teste de Grubbs (1969), para normalidade, tem estatística $G = \text{máx } |X_i - X|/s$, em que X_i seria aberrante e X e *s* são, respectivamente, a média e o desvio padrão amostrais obtidos na ausência do pretenso valor aberrante.

Conforme Brownfield & Kalivas (2017), esses testes, como do T-Student, q-resíduos e resíduos de modelos, são padrão em química analítica. Esses autores extendem testes de somas de diferenças de postos para esses valores no espectro e no analito utillizando análise de procrustes medições de espectroscopia no infravermelho próximo (NIR).

O teste de Hampel (2001), mais atual, não tradicional e não paramétrico, é pouco conhecido. Define-se $r_i = (x_i - M_e)$, em que x_i são as observações (medições), i = 1, 2, ..., n e M_e é a mediana das observações; em seguida, se obtém a mediana do módulo de r_i, quer seja, $Me|r_i|$, e se admite x_i como aberrante quando $|r_i| \geq 4{,}5 \times Me|r_i|$.

Testes menos comuns, como o h de Mandel em *American Society for Testing Materials* (2018), para médias de laboratórios, também vêm sendo usados em interlaboratoriais e escores Z = (valor amostral – média populacional)/(desvio padrão populacional) são comumente usados para classificar desempenho de laboratórios.

A pesquisa em gráficos estatísticos teve grande impulso na década de 1980, brevemente se destaca Cleveland & McGill (1985), além de Wilk & Gnanadesikan (1968), que mostraram gráficos distribucionais, de que também se podem extrair valores aberrantes. Tukey (1977) elaborou desenhos esquemáticos (*boxplots*) para visualizar massas de dados, intervalos de variação, valores aberrantes, etc. Bartolucci et al. (2015) trazem alguns gráficos e técnicas para esses valores em laboratórios, sem discutir os resultados.

Assim, como discussão de resultados dos testes não foi encontrada em alimentos, buscou-se ampliar a identificação desses valores usando alguns gráficos descritivos e exploratórios, confrontá-los com testes T- Student, Dixon, Grubbs e Hampel, bem como comparar os últimos e discutir seus resultados. Para maior compreensão do trabalho, foram empregados interlaboratoriais publicados e se traçou uma discussão dos resultados.

2 Material e métodos

Principais materiais foram quatro experimentos interlaboratoriais publicados, cujas informações a seguir constituem praticamente tudo o que se encontra disponível, quer sejam:

a) um experimento de precisão de Youden & Steiner (1975) apud Hamaker (1986) para uma amostra homogênea cujo teor de lipídeos (%) foi medido em 11 laboratórios, com duas repetições/laboratório;

b) outro experimento de Youden & Steiner (1975) apud Hamaker (1986) para determinar o teor de umidade (%) com 10 laboratórios em três níveis em parcelas subdivididas, em que o material foi distribuído aos laboratórios em duplicatas/laboratório;

c) um experimento de precisão, também chamado estudo colaborativo, em parcelas subdivididas de matéria insaponificável (%) para três amostras publicado na *International Organization for Standardization* (1988), realizado com 14 laboratórios em duplicatas/laboratório, que também teve seus resíduos estudados conforme o modelo do delineamento experimental desse tipo de experimento; e

d) o experimento para N-total (%) em porção homogeneizada de linguiças, posteriormente embalada, em oito laboratórios, analisando-se dois pacotes/laboratório de Suhre et al. (1982).

Foram elaborados desenhos esquemáticos e diagramas de ramo-e-folhas (Tukey, 1977), bem como gráficos de probabilidade normal e de dispersão elaborados no *Statgraphics Plus* (Manugistics, 1993). Cálculos estatísticos foram realizados no *Statistical Analysis System Institute* (1985), com obtenção de escore Z robusto (Z_r), derivados de Maronna et al. (2006), ajustes de modelos de delineamento e regressões (Drapper & Smith, 1981), além de cálculos de probabilidades e testes de hipóteses.

Foram realizados os testes de Grubbs (1969), Dixon (1950) e Hampel (2001), o último não referenciado em alimentos. O teste t-Student foi usado como indicativo e níveis de significância de testes giraram em torno de 5% e, mencionado o delineamento do experimento de precisão, foram examinados seus resíduos.

Similarmente a ensaios de proficiência, para $|Z_r| \geq 3$, a medição foi considerada aberrante, para $2 < |Z_r| < 3$, questionável, e se $0 \leq |Z_r| \leq 2$, a medição não seria aberrante.

3 Resultados e discussão

3.1 Experimento de precisão com uma única amostra

No experimento de Youden & Steiner (1975) apud Hamaker (1986) para lipídeos (%), *não* foi detectado valor aberrante nas Figuras 1a, b, c, o que consta na Tabela 1. As probabilidades do maior e do menor resultado nesse experimento (Tabela 2), apontaram t-Student significativo para 22,0, indicando-o como aberrante potencial. No entanto, os testes de Dixon e Grubbs não o corroboraram e fato similar ocorreu com Zr e teste de Hampel.

Desse modo, tanto graficamente (Figura 1b, c), quanto com uso de testes de hipóteses, concluiu-se pela inexistência de valor aberrante nesse experimento, conforme publicado por Youden & Steiner (1975) apud Hamaker (1986).

Tabela 1. Valores aberrantes identificados em análises gráficas segundo experimento pesquisado.

Fonte	Amostra ou nível	PAGPN	C	Desenho Esquemático	Ramo-e-folhas	Aberrante Publicado	Concordância
1	único	22,0	não	nenhum	nenhum	nenhum	sim
2	1	nenhum	sim	nenhum	nenhum	nenhum	sim
	2	nenhum	sim	nenhum	nenhum	nenhum	sim
	3	22,0	não	22,0	22,0	nenhum	não
3	1	diversos	sim	0,78; 0,89	0,78; 0,89	0,78; 0,89	sim
	2	nenhum	não	0,51	0,51	nenhum	não
	3	diversos	não	0,40; 0,41	0,40; 0,41	0,40; 0,41	sim
	resíduos	diversos	não	0,16; 0,18; 0,23; 0,24; 0,27; 0,27	0,16; 0,18; 0,23; 0,24; 0,27; 0,27	nenhum	não
4	único	nenhum	sim	9,79	9,79	9,79; 11,23	não

PAGPN: possível valor aberrante conforme o gráfico probabilidade normal. C: existência de resultados coincidentes no mesmo ponto (não ou sim). Concordância: concordância de resultados de análises gráficas.

1. Youden & Steiner (1975) apud Hamaker (1986), lipídeos (%) em carne medidos em 11 laboratórios, com duas repetições/laboratório;
2. Youden & Steiner (1975) apud Hamaker (1986), teor de umidade (%) medido em 10 laboratórios em três níveis, parcelas subdivididas com material distribuído em duas porções/laboratório;

3. Interlaboratorial do tipo experimento de precisão, matéria insaponificável (%) em três amostras da *International Organization for Standardization* (1988) com 14 laboratórios e duplicatas/laboratório;
4. N-total (%) em porção homogeneizada de linguiças, em oito laboratórios, dois pacotes/laboratório (Suhre et al., 1982).

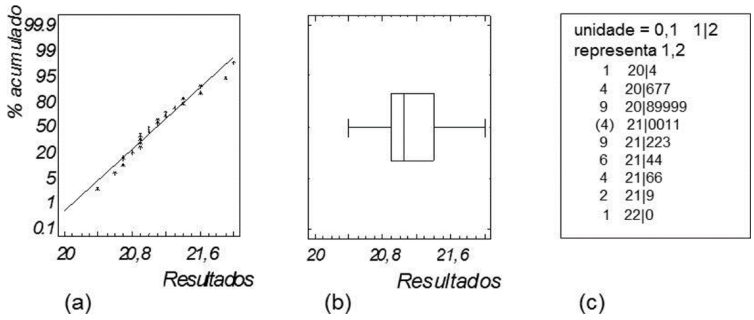

Figura 1 - Gráfico de probabilidade normal (a), respectivo desenho esquemático (b) e diagrama de ramo-e-folhas (c) dos resultados de Youden & Steiner (1975) apud Hamaker (1986) para teor de lipídeos (%).

3.2 Experimento em parcela subdivididas do tipo de Youden

A Figura 2, com três níveis de umidade, não demonstrou valores aberrantes para o primeiro nível nas Figuras 2a, b, c. Fato similar foi notado no segundo nível (Figuras 2d, f). Porém, a Figura 2g indicou que o maior valor do terceiro nível estaria desviado dos demais, o desenho na Figura 2h o mostrou como aberrante e o diagrama (i) *ratificou* essa indicação.

Portanto, graficamente, teria sido detectado o valor aberrante 22 no nível 3. Adicionalmente, esse valor seria aberrante também segundo os testes T-Sudent, de Grubbs e de Hampel, além do escore Z_r (Tabela 2), discordante do publicado.

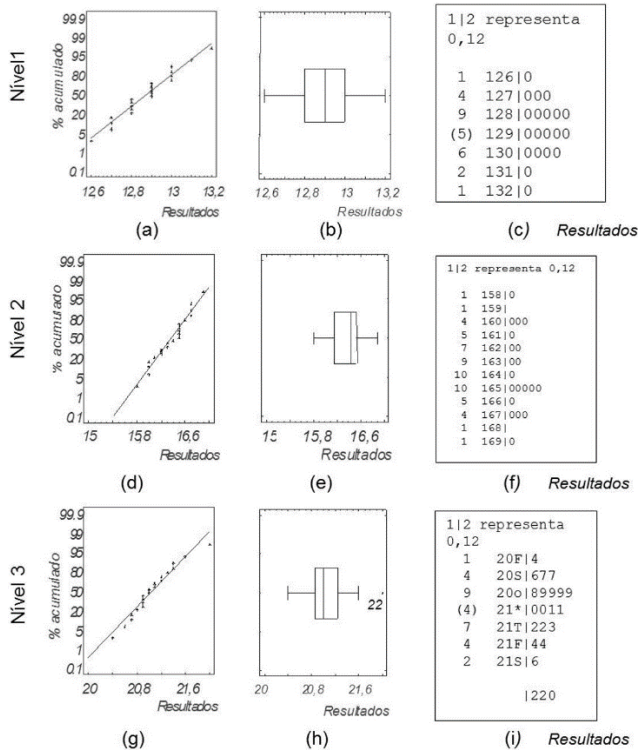

Figura 2. Gráfico de probabilidade normal, desenho esquemático e diagramas de ramo-e-folhas dos resultados do experimento em parcelas subdivididas de umidade em porcentagem (%) para o nível 1 (a, b, c), nível 2 (d, e, f) e nível 3 (g, h, i).

Nos níveis 2 e 3 da Figura 2, foram notados indícios de não normalidade, o que pode ter interferido nos resultados da publicação mencionada. Desse modo, não foi obtida conformidade com os resultados publicados para esse interlaboratorial.

3.3 Experimento de matéria insaponificável em três níveis

A Figura 3a, de matéria insaponificável (%) da *International Organization for Standardization* (1988), mostrou diversos candidatos a valores aberrantes nos extremos da escala da amostra 1 e, nessa amostra, foram registrados três valores aberrantes "simultâneos" no desenho (b), quer sejam, 0,78 e 0,89 (duas vezes), correspondentes aos laboratórios 6 e 9, que foram também classificados como aberrantes no diagrama (c).

Por outro lado, para a amostra 1 (Tabela 2), os testes de Dixon e Grubbs, além de Z_r, indicaram 0,78 como aberrante e esses mesmos testes, mais o de Hampel, também indicaram os dois valores 0,89 como aberrantes. Assim, os gráficos usados foram sensíveis ao apontar 0,78 e 0,89 como aberrantes.

Segundo o T-Student, o valor 0,51 na amostra 2 seria aberrante, Z_r trouxe 0,18 como aberrante, enquanto que o teste de Hampel admitiu somente o primeiro, 0,51. Note-se que 0,51 não foi indicado nessa amostra pelos testes tradicionalmente usados.

Finalmente, segundo a Figura 3g, h e i, os valores que poderiam ser aberrantes na amostra 3 foram vários: 0,03; 0,04; 0,40; e 0,41. Porém, nem todos foram confirmados pelos três gráficos conjuntamente. Os valores 0,03 e 0,04 estão muito próximos, o que não confirmaria algum deles como aberrante. Adicionalmente, o desenho esquemático também não os indicou e a Tabela 2 não os confirmou.

A *International Organization for Standardization* (1988) traz os dois valores reconhecidos como aberrantes citados para a amostra 1, um segundo o teste de Cochran e outro pelo de Dixon. Porém, na amostra 2, segundo o autor original, nenhum laboratório teria apresentado valores aberrantes e o laboratório 11, na amostra 3, foi eliminado pelo teste de Dixon.

O valor 0,40 foi classificado como aberrante nos três testes tradicionais usados (Tabela 2) e também indicado pelo escore Zr, porém não confirmado pelo teste de Hampel. Complementarmente, o valor 0,41 foi indicado pelos testes de Dixon e Grubbs. Já os valores 0,03 e 0,04 não foram indicados como aberrantes pelos testes na Tabela 2. Escores Zr obtidos nesse experimento serão discutidos no item 3.7.

Assim, este artigo divergiu ao identificar os valores aberrantes na amostra 1 e a *International Organization for Standardization* (1988) não identificou o valor aberrante 0,51 na amostra 2.

Os resultados da Figura 3h, i 0,40 e 0,41 tenderam a indicar vícios comumente procedentes de experimentos não cegos, que frequentemente causam desvios em resultados laboratoriais.

Aprofundando a verificação com o modelo de delineamento, a Figura 4a teria demonstrado diversos valores aberrantes potenciais enquanto que a Figura 4b apontou 5 (cinco), correspondentes às medições 0,78; 0,51; 0,4; 0,41; e 0,89, nessa ordem. Nos testes de hipóteses para esses cinco valores (Tabela 2), o valor 0,78 não foi confirmado. No entanto, o escore Z_r e o teste de Hampel indicaram quatro valores potencialmente aberrantes nas amostras 1, 2 e 3.

Porém, considerando o modelo estocástico do experimento, foram admitidos como aberrantes potenciais 0,40; 0,41; 0,51; e 0,89, não conformes com os publicados para as amostras 1, 2 e 3. As indicações da Figura 3 também trouxeram conclusões próximas disso. Dessa forma, os desenhos da Figura 4b mostraram boa detecção de valores aberrantes quando se considerou o modelo que originou as medições nos cálculos.

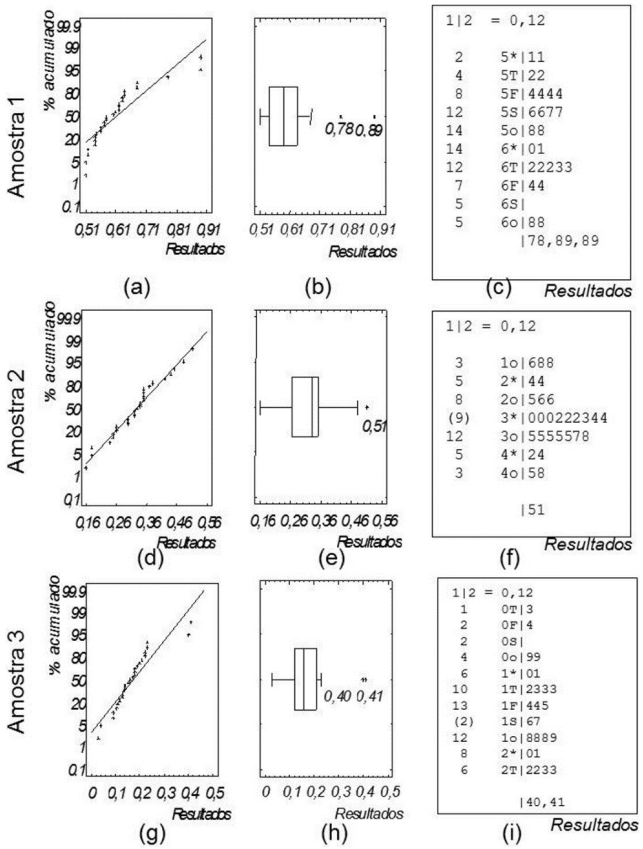

Figura 3. Gráfico de probabilidade normal, desenhos esquemáticos e diagrama de ramo-e-folhas dos resultados de matéria insaponificável para a amostra1 (a, b, c), amostra 2 (d, e, f) e amostra 3 (g, h, i).

Tabela 2. Valores detectados como possíveis aberrantes, resultados dos respectivos testes de hipóteses e escore Z_r segundo experimento revisado.

Fonte	Amostra ou Nível	Valor Investigado	Nº	Dixon	Grubbs	t-Student	Z_r	Hampel*
1	único	20,4	22	0,13	0,01	< 0,00	-0,97	nenhum
		22,0		0,06	2,52	2,16 *	1,19	
2	1	12,6	20	0,17	2,11	-1,85	-1,48	nenhum
		13,2		0,17	2,62	2,19 *	1,75	
	2	15,8	20	0,18	2,05	1,81	*2,48*	nenhum
		16,9		0,18	2,25	1,95	0,58	
	3	20,4	20	0,25	2,00	1,77	*-2,59*	
		22,0		0,13	3,29 *	2,56 *	**3,75**	22,00
3	1	0,51	28	0,03	1,13	-1,05	**-3,69**	
		0,78		0,37 +	3,88 *	1,69	**5,93**	
		0,89		0,29 *	4,73 *	2,80	**9,85**	2 x 0,89
	2	0,18	28	0,18	2,31	-1,89	**-4,41**	
		0,51		0,09	2,50	2,22 *	**5,18**	0,51
	3	0,03	28	0,03	1,78	-1,66	*-2,87*	
		0,40		0,46 *	4,60 *	2,73 *	**4,73**	0,40
		0,41		0,03	3,50 *	2,85 *	**4,93**	0,41
	resíduos	0,03	85	0,01	-0,30	-1,86	-0,65	
		0,40		0,10	2,71	2,58	**5,84**	0,40
		0,41		0,02	2,84	2,69	**6,09**	0,41
		0,51		0,04	2,14	2,07	**4,67**	
		0,78		0,03	1,92	1,87	**4,23**	
		2 x 0,89		0,08	3,60 *	3,10 *	**7,02**	2 x: 0,89

Tabela 2. Continuação...

Fonte	Amostra ou Nível	Valor Investigado	Nº	Dixon	Grubbs	t-Student	Z_r	Hampel*
		9,79		0,14	2,64	-2,49 *	*-2,45*	
4	único	11,23	64	0,11	2,12	2,03	*2,01*	nenhum
		11,33		0,03	2,48	2,35 *	*2,31*	

1: Youden & Steiner (1975) apud Hamaker (1986), de lipídeos (%) em carne medidos em 11 laboratórios, com duas repetições/laboratório. 2: Youden & Steiner (1975) apud Hamaker (1986), teor de umidade (%) medido em 10 laboratórios em três níveis, parcelas subdivididas com material distribuído em duas porções/laboratório. 3: interlaboratorial do tipo experimento de precisão, matéria insaponificável (%), três amostras (*International Organization for Standardization*, 1988) com 14 laboratórios e duplicatas/laboratório. 4: N-total (%) em porção homogeneizada de linguiças, em oito laboratórios, dois pacotes/laboratório (Suhre et al., 1982). Hampel*: valor aberrante significativo $p \leq 0,05$.

3.4 Experimento de N-total (%) em linguiças

Suhre et al. (1982) encontraram como valores aberrantes 9,79 e 11,23. A Figura 5b confirmou o primeiro, mas não mostrou valores ao redor de 11,3% como aberrantes. Complementarmente, a Tabela 2 apresentou t-Student significativos para esses dois pontos, porém nenhum dos outros testes ou indicadores da Tabela obteve o mesmo resultado.

Desse modo, concluiu-se que não houve valores aberrantes nesse experimento, assim houve não conformidade quanto aos três valores superiores da distribuição que constituiriam valores aberrantes segundo a publicação original.

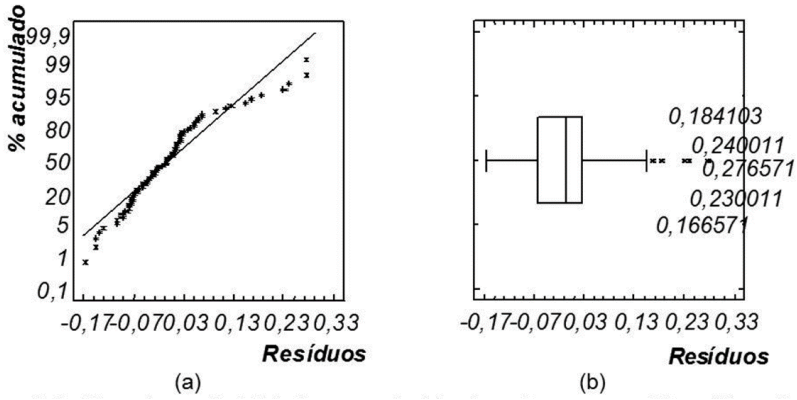

Figura 4. Gráfico de probabilidade normal (a), desenho esquemático (b) e diagrama de ramo-e-folhas dos resíduos do experimento de matéria insaponificável em parcelas subdivididas.

3.5 Gráficos e resultados publicados baseados em testes

A Tabela 1 compara os métodos gráficos empregados e os resultados obtidos na literatura consultada. Nela, há de se notarem discordâncias com resultados publicados com base em testes de hipóteses. Na competição entre os gráficos e os testes empregados, os testes de hipóteses perderam quanto ao número de valores aberrantes detectados.

Os gráficos estudados foram mais sensíveis que diversos testes de hipóteses, na medida em que retratam a distribuição das medições, sua forma e seu comportamento, destacando prováveis aberrantes. Assim, embora desenhos esquemáticos sejam mais comumente usados que os diagramas de ramo-e-folhas, os últimos pareceram mais eficientes nessa identificação frente ao que é exposto no exame da distribuição das medições.

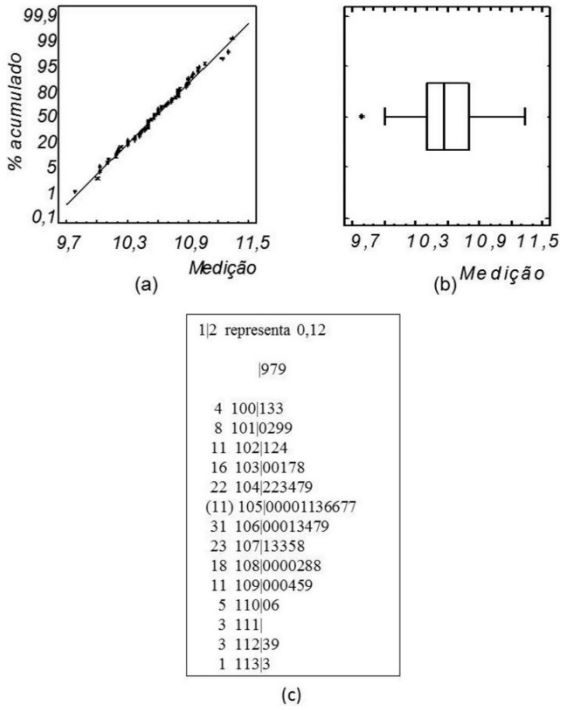

Figura 5. Gráfico de probabilidade normal (a), desenho esquemático (b) e diagrama de ramo-e-folhas (c) para nitrogênio total (%) em linguiças em Suhre et al. (1982).

3.6 Concordância entre testes de hipóteses e outros detalhes

A Tabela 2 traz uma comparação dos resultados de alguns testes para valores aberrantes comumente aplicados em métodos analíticos em alimentos.

Os testes de Grubbs e t-Student são matematicamente similares, mas o t-Student é mais vantajoso que o de Grubbs nos casos normais, especialmente quando há quantidade expressiva de medições sendo analisadas. Assim, o teste t-Student qualificou como aberrante o valor 13,2 no primeiro nível do experimento em parcelas subdivididas (Tabela 2) e, de mesmo modo, os valores 9,79 e 11,33 no experimento de N-Total (%) em linguiças.

Ainda, comparando os testes de Grubbs e t-Student, o primeiro não indicou como aberrantes diversos dos valores indicados pelo segundo. Desse modo, o teste de Grubbs demonstrou baixo grau de detecção de valores aberrantes frente ao segundo, o que pode ser explicado pelo baixo poder desse teste. O teste de Dixon raramente foi concordante com os demais testes, além de indicar a mais baixa quantidade de valores potencialmente aberrantes.

Já o teste de Hampel, por ser não paramétrico e ter distribuição exata, seria o mais aplicável, e dispensa normalidade, que é exigida por esses três. Ainda, não demonstrou grande disparidade quanto aos gráficos e apresentou probabilidade de valores aberrantes compatível com a esperada em vários experimentos. Portanto, esse teste obteve o melhor desempenho nos casos pesquisados.

Porém, as formas com que medições (quaisquer dados) podem ser visualizadas ou analisadas são diversas e não há uma forma única ou análise recomendável com esse fim. Considerar os modelos experimentais de delineamento ou regressão, em que repousam as medições, é uma forma recomendável de análise.

Não se pode admitir um número de aberrantes como trazem *Association of Official Analytical Chemists* (1995, 2002).

Identificar valores aberrantes em 10 a 20 medições não é simples e pode ferir os padrões estocásticos uma vez que as probabilidades de erro comum ou grave, geralmente admitidas em alimentos, são por volta de 5 ou 1%. Contudo, de uma em vinte a uma em dez, ou seja, haver 5 a 10% de medições aberrantes compõe ou

ultrapassa 5% das medições disponíveis e, então, haveria probabilidades de aberrantes superiores às probabilidades de erro.

Adicionalmente, a expressão "raramente obtidos" da *Food and Drug Administration* (2006) pode ser equivocada, pois deve-se lembrar que a probabilidade desses valores, embora baixa, não é nula, ou seja, eles podem ocorrer sem que constituam erros de medição.

Pelo exposto e pelos resultados obtidos, o escore Z_r indicou a maior quantidade de valores candidatos a aberrantes, mas a probabilidade com que o fez nos experimentos estudados, chegando a 10% na amostra 1 do experimento de matéria insaponificável, obriga a admiti-lo como um indicador que pode gerar falsos positivos. Por outro lado, deixar de empregar um escore como esse, ou similar, pode resultar desprezar possíveis desvios de medição, constatados posteriormente à sua realização.

Assim, não se recomenda excluir valores aberrantes com probabilidade maior que 1% sem verificar detidamente sua causa, e o teste de Hampel, recente, seria mais adequado que os testes de Dixon, T-Student e Grubbs.

4 Conclusões

O gráfico que melhor indicou candidatos a aberrantes foi o diagrama de ramo-e-folhas, que não gerou resultados divergentes quando comparado com o desenho esquemático, enquanto que o de probabilidade normal foi o menos eficiente na indicação de valores aberrantes posteriormente confirmados.

Youden & Steiner (1975) apud Hamaker (1986), para uma amostra, e esta pesquisa obtiveram os mesmos resultados. Assim, não houve valores aberrantes segundo as duas pesquisas.

No entanto, os resultados para valores aberrantes nos demais experimentos não se mostraram conformes com as publicações que os geraram. No terceiro nível do experimento para teor de umidade em Hamaker (1986), um valor comprovadamente aberrante não foi acusado por esse autor. Também se divergiu ao identificar os valores aberrantes no experimento de matéria insaponificável. Finalmente, na matriz estudada de Suhre et al. (1982), nenhum valor de N-total (%) foi presentemente detectado como aberrante, contrariamente ao anteriormente obtido por esses autores.

Tecnicamente, dentre os testes probabilísticos, o de Hampel seria indicado nesses casos. Esse teste se mostrou adequado para medições em laboratórios de alimentos por não exigir normalidade e pelas probabilidades dos valores identificados encontradas.

O escore Z_r, outro meio numérico para identificar esses valores, também não exige normalidade e foi um indicador detalhista de valores aberrantes potenciais. Como há certa probabilidade de falsos positivos, pode ser usado como um alerta.

As principais noções não mudaram, os métodos tradicionais continuam sendo úteis e se espera que o teste de Hampel, mais adequado que os tradicionais, e os métodos gráficos sejam rotineiramente empregados.

Valores aberrantes em potencial devem ser verificados exaustivamente e não se sugere sua exclusão, especialmente, quando apresentem incidência maior que 1%. Finalmente, ao se adotarem métodos de medição multidimensionais (ou multivariados), a identificação de valores aberrantes unidimensionais (univariados) deve ser isoladamente realizada.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pelo suporte financeiro.

Referências

Association of Official Analytical Chemists – AOAC. (1995). Guidelines for collaborative study procedures to validate characteristics of a method of analysis. *Journal of the American Oil Association Chemists*, *78*(1), 143A-160A.

Association of Official Analytical Chemists – AOAC. (2002). *Appendix D: Guidelines for collaborative study procedures to validate characteristics of a method of analysis*. Recuperado em 3 de março de 2020, de http://members.aoac.org/aoac_prod_imis/AOAC_Docs/StandardsDevelopment/Collaborative_Study_Validation_Guidelines.pdf

American Society for Testing Materials – ASTM. (2018). *ASTM D4483-18: Standard practice for evaluating precision for test method standards in the rubber and carbon black manufacturing industries.* Recuperado em 3 de março de 2020, de https://reference.globalspec.com/standard/4638645/astm-d4483-18

Bartolucci, A., Singh, K. P., & Bae, S. (2015). *Methodologies in outlier analysis*. New York: Wiley.

Brownfield, B., & Kalivas, J. H. (2017). Consensus outlier detection using sum of ranking differences of common and new outlier measures without tuning parameter selections. *Analytical Chemistry*, *89*(9), 5087-5094. PMid:28367620. http://dx.doi.org/10.1021/acs.analchem.7b00637

Chauvenet, W. (1863). *A manual of spherical and practical astronomy* (Vol. II). London: J.B. Lippincott.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828-833. PMid:17777913. http://dx.doi.org/10.1126/science.229.4716.828

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Human Genetics*, *11*(1), 47-52.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics*, *21*(4), 488-506. http://dx.doi.org/10.1214/aoms/1177729747

Drapper, N. R., & Smith, H. (1981). *Applied regression analysis.* New York: Wiley.

Dunn, O. J., & Clark, V. A. (1987). *Applied statistics: Analysis of variance and regression.* New York: Wiley.

Food and Drug Administration – FDA. (2006). *Guidance for industry, investigating out-of-specification (OOS) test results for pharmaceutical production.* Silver Spring: FDA.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21. http://dx.doi.org/10.1080/00401706.1969.10490657

Hamaker, H. C. (1986). A statistician's approach of repeatability and reproducibility. *Journal of the American Oil Association Chemists*, *69*(3), 417-428. http://dx.doi.org/10.1093/jaoac/69.3.417

Hampel, F. (2001). Robust statistics: A brief introduction and overview. In: Eidgenossische Technische Hochschule. *Robust statistics and fuzzy techniques in geodesy and GIS.* Zurich: ETH Zürich.

Horwitz, W. (1995). Protocol for the design, conduct and interpretation of method-performance studies. *Pure and Applied Chemistry*, *67*(2), 331-343. http://dx.doi.org/10.1351/pac199567020331

Instituto Nacional de Metrologia, Qualidade e Tecnologia – INMETRO. (2016). *Orientações sobre validação de métodos de ensaios químicos, rev. 5 (DOQ-CGRE-008).* Rio de Janeiro: INMETRO.

International Organization for Standardization – ISO. International Eletrotechnical Comission – IEC. (2005). *General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025).* Genève: ISO.

International Organization for Standardization – ISO. International Eletrotechnical Comission – IEC. (1988). *ISO 3596-2: Animal and vegetable fats and oils: Determination of unsaponifiable matter: Part 2: Rapid method using hexane extraction.* Genève: ISO.

Manugistics. (1993). *Statgraphics reference manual.* Cambridge: Manugistics.

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods.* New York: Wiley. http://dx.doi.org/10.1002/0470010940.

Molenaar, J., Cofino, W. P., & Torfs, P. J. J. F. (2018). Efficient and robust analysis of interlaboratory studies. *Chemometrics and Intelligent Laboratory Systems*, *175*(4), 65-73. http://dx.doi.org/10.1016/j.chemolab.2018.01.003

National Institute of Standards and Technology – NIST. Semiconductor Manufacturing Technology. (2020). *NIST/SEMATECH e-handbook of statistical methods.* Recuperado em 10 de março de 2020, de https://www.itl.nist.gov/div898/handbook

Organisation for Economic Co-operation and Development – OECD. (2018). *OECD guideline for the testing of chemicals, 442B.* Paris: OECD.

Pang, G., & Cao, L. (2020). Heterogeneous univariate outlier ensembles in multidimensional data. *ACM Transactions on Knowledge Discovery from Data*, *14*(6), 1-27. http://dx.doi.org/10.1145/3403934

Peirce, B. (1852). Criterion for the rejection of doubtful observations. *The Astronomical Journal*, *2*(45), 161-163. http://dx.doi.org/10.1086/100259

Pharmceutical Technology Editors. (2006). A review of statistical outlier methods. *Pharmaceutical Technology*, *30*(11), 82-86.

Seaman, J. E., & Allen, E. (2010). Outlier options: Consider simple parametric tests to find an outlier's significance. *Quality Progress*, *43*(1), 56-57.

Statistical Analysis System Institute – SAS Institute. (1985). *SAS user's guide: Statistics* (Vol. 5). Cary: SAS Institute.

Suhre, F. B., Corrao, P. A., Glover, A., Malanoski, A. J., Cannon, L. D., Dummett, T., Funk, R., Glover, A., Heavner, G., Hoover, R. L., Latham, M., Long, F. L., Martini, J. H., McGee, K., Morris, W. C., Oberste, W., Okamoto, M., Pakrasi, B., Pasquarella, P. J., Reiser, J., Sorensen, L., Lovestrand, J., Taylor, M., Trombella, B., Warden, S. R., Wayo, C., Wiebke, R., & Woods, W. (1982). Comparison of three methods for determination of crude protein in meat: Collaborative study. *Journal of the Association of Analitycal Chemists*, *65*(6), 1339-1345. PMid:7174576. http://dx.doi.org/10.1093/jaoac/65.6.1339

Tukey, J. W. (1977). *Exploratory data analysis* (3rd ed). Reading: Addison-Wesley.

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*(1), 1-17. PMid:5661047. http://dx.doi.org/10.2307/2334448

Youden, W. J., & Steiner, E. H. 1975. *Statistical manual of the AOAC.* Arlington: AOAC International.

Zhongya, F., Feng, H., Jiang, J., Zhao, C., Jiang, N., Wang, W., & Zeng, F. (2020). Monte Carlo optimization for sliding window size in Dixon quality control of environmental monitoring time series data. *Applied Sciences*, *10*(1), 1-14.

Financiamento: Nenhum.

Received: Dec. 02, 2020; Accepted: May 12, 2021